*Healthcare Transformation with Informatics and Artificial Intelligence*
*J. Mantas et al. (Eds.)*

*doi:10.3233/SHTI230554*

# Mining Greek Tweets on Long COVID Using Sentiment Analysis and Topic Modeling

Afroditi KATIKA[a,b,1], Emmanouil ZOULIAS[a], Vassiliki KOUFI[b] and Flora MALAMATENIOU[a]
[a] *Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece*
[b] *Department of Digital Systems, School of Information and Communication Technologies, University of Piraeus, Piraeus, Greece*

**Abstract.** Around 10% to 20% of patients experience Long COVID after recovering from COVID-19. Many people are turning to social networks such as Facebook, WhatsApp, Twitter, etc., to express their opinions and feelings regarding Long COVID. In this paper, we analyse text messages in the Greek language posted on the Twitter platform in 2022 to extract popular discussion topics and classify the sentiment of Greek citizens regarding Long COVID. Results highlighted the following discussion topics: Greek-speaking users discuss Long COVID effects and time required to heal, Long COVID effects in specific population groups like children and COVID-19 vaccines. 59% of analysed tweets conveyed a negative sentiment while the rest had positive or neutral sentiment. The analysis shows that public bodies could benefit from systematically mining knowledge from social media to understand public's perception of a new disease and take action.

**Keywords.** Big Data, Twitter, Natural Language Processing, NLP, Sentiment Analysis, COVID-19, Pandemic

## 1. Introduction

Long COVID syndrome is defined by the World Health Organization (WHO) as the continuation or development of new symptoms three months after the initial COVID-19 infection, with these symptoms lasting for at least two months with no other explanation [1]. In Greece, a questionnaire was filled by patients that suffered from Long COVID and 66.8% claimed that their symptoms linger for more than six months [2].

Social Media platforms like Twitter are widely used by citizens who express their opinions and reactions and therefore constitute a suitable source to assess what the population is talking about in real-time [3]. In order to systemically monitor what people are saying, natural language processing is used. Topic Modelling and Sentiment Analysis are methods that have been used extensively in literature to mine knowledge from social media especially in the advent of COVID-19 [4, 5].

In this paper, a study is presented which analyses posts on Twitter, regarding Long COVID to better understand the opinions and sentiments of the Greek public. The

[1] Corresponding Author: Afroditi KATIKA, Athens, Greece; E-mail: afro.katika@gmail.com.

objective of this study is two-fold: a) to extract most discussed topics about Long COVID on Twitter in Modern Greek and b) to categorise the sentiment of those tweets.

## 2. Methods

Tweets were extracted through Twitter's API using Python. In order to construct the dataset the following filters were applied: tweets were posted between January 1st and 31st December 2022, in Greek language and contained either the words '*Long COVID*' or *#longCOVID*. 1000 of all tweets were saved in a dataset which, in turn, was preprocessed. Accents were removed and all words were transformed to lower case so that they could be easily compared to each other. Then, the following words were removed:

- Links, usernames and words that were used for extracting the dataset
- Greek stopwords combining nltk stopwords [6] and stopwords-iso lists [7]
- Words that consisted of three or less characters
- Greek words for COVID ('κορωνοιός') and pandemic ('πανδημία΄) as well as other ways these words are written
- Tweets that were deemed duplicate

A wordcloud as shown in Figure 1 was produced to highlight most common words.

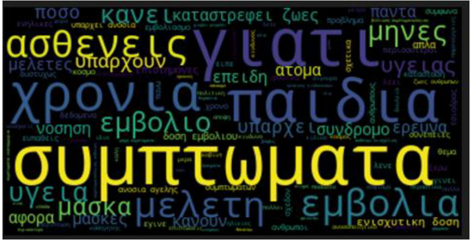


**Figure 1.** Wordcloud for the dataset

An important step to reduce further the variations of the same word is lemmatisation. To this end, spacy pipeline was used [8].

Two different topic modelling methods were employed to reveal discussion topics. The first one was Latent Dirichlet Allocations (LDA) through Gensim's open-source library. After tokenising tweets, a dictionary was created. The dictionary included all the words apart from those that appeared in less than two tweets or appeared in more than 99% of the corpus. In order to identify best number of topics, all numbers of topics between 2 and 19 were tried and their performance was evaluated based on the coherence score. As shown in Figure 2 the coherence score varied between 0.43-0.52.

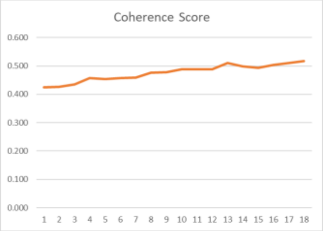


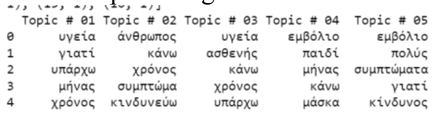
**Figure 2.** Coherence score per number of topics

The second method used was Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) which is more suited to shorter texts and more specifically, the rwalk [9].

Regarding sentiment analysis, less pre-processing was required to avoid altering the sentiment of the sentence. On the original dataset, usernames, links, accents and duplicate tweets were removed and all words were lower-cased, leaving the dataset with 937 tweets. Greek-BERT was used which has been trained solely in Greek texts [10]. To overcome the challenge of lack of a training set, a dataset that is available online and classifies the sentiment of e-commerce reviews[11] into negative and positive was used. This restricted the range of sentiments that could be used but created a training dataset with 6552 reviews. Model achieved an approximated accuracy of 94% compared to a human annotator's classifications on our Long COVID dataset as shown in Table 1.

## 3. Results

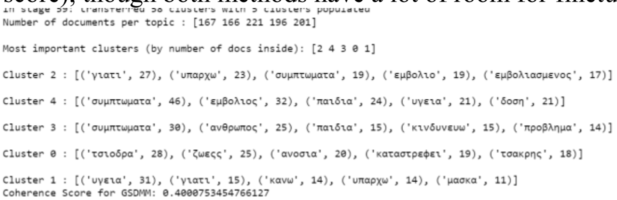
Following pre-processing, a wordcloud was created on the dataset (Figure 1). Most popular words were: *symptoms, children, years, why* and *patients*. Followed by: *study, vaccines, months* and *mask*. Since the coherence score for the LDA implementation did not vary significantly between the numbers of topics we assessed, 5 topics were selected (with a coherence score of 0.45) so that they could be easier assessed and apprehended by a human. As shown in Figure 3, for each topic, 5 most popular words were printed. First topic is concerned with health and time required to heal after Long COVID. The second topic concerns Long COVID symptoms. Third topic is similar to the first one. Fourth topic concerns children and applying masks and the fifth topic discusses COVID-19 vaccines and their relationship to Long COVID.



```
  Topic # 01 Topic # 02 Topic # 03 Topic # 04  Topic # 05
0      υγεία   άνθρωπος      υγεία   εμβόλιο     εμβόλιο
1      γιατί       κάνω    ασθενής     παιδί       πολύς
2     υπάρχω     χρόνος       κάνω     μήνας  συμπτώματα
3      μήνας   συμπτώμα     χρόνος      κάνω       γιατί
4     χρόνος  κινδυνεύω     υπάρχω     μάσκα     κίνδυνος
```

**Figure 3.** LDA topics

For a direct comparison, 5 topics were also pre-defined in GSDMM, as shown in Figure 4, which achieved a coherence score of 0.40. Topic results were quite similar to the ones produced by LDA. Overall, LDA showed a better performance (evaluated on the coherence score), though both methods have a lot of room for finetuning.



```
In stage 59: transferred 58 clusters with 5 clusters populated
Number of documents per topic : [167 166 221 196 201]

Most important clusters (by number of docs inside): [2 4 3 0 1]

Cluster 2 : [('γιατι', 27), ('υπαρχω', 23), ('συμπτωματα', 19), ('εμβολιο', 19), ('εμβολιασμενος', 17)]

Cluster 4 : [('συμπτωματα', 46), ('εμβολιος', 32), ('παιδια', 24), ('υγεια', 21), ('δοση', 21)]

Cluster 3 : [('συμπτωματα', 30), ('ανθρωπος', 25), ('παιδια', 15), ('κινδυνευω', 15), ('προβλημα', 14)]

Cluster 0 : [('τσιοδρα', 28), ('ζωεςς', 25), ('ανοσια', 20), ('καταστρεφει', 19), ('τσακρης', 18)]

Cluster 1 : [('υγεια', 31), ('γιατι', 15), ('κανω', 14), ('υπαρχω', 14), ('μασκα', 11)]
Coherence Score for GSDMM: 0.4000753454766127
```

**Figure 4.** GSDMM topics

Greek-BERT was applied for tweets sentiment analysis. Out of the 937 tweets, 555 had a negative sentiment while 382 positive. A sample of the tweets (457) was annotated by a human. The comparison, as shown in Table 1, showed a model accuracy of 94% (6% of the sample was mis-classified). Human inspection revealed some of the positive tweets could be re-classified as neutral if more categories were applied in training.

**Table 1.** Human classification versus Greek-BERT on a sample of the dataset

| Results(Human/BERT) | Count | Percentage |
|---|---|---|
| Positive/Positive | 169 | 0.370 |
| Positive/Negative | 2 | 0.004 |
| Negative/Negative | 261 | 0.571 |
| Negative/Positive | 25 | 0.055 |

## 4. Discussion and Conclusions

In this paper, we examined sentiments and discovered key topics in Long COVID-related messages posted by Twitter users in Greece for the year 2022. Topic Modelling revealed approximately 5 topics (but 4 discreet ones to the human eye), with no special differentiation between LDA and GSDMM, even though LDA performed better on coherence score. With regard to the sentiments produced by the model, 59.2% of the tweets carried a negative sentiment and 40.7% carried a positive or neutral sentiment showing that public bodies could benefit of such kind of models to better understand public's perception of a new disease and take action where needed. Our results in terms of percentages are similar to an earlier study on Greek tweets that analysed reactions to the COVID-19 vaccines [4]. Further studies in this area could investigate a bigger range of emotions as well as topics in specific emotions so that actions could be more targeted.

## References

[1] Post COVID-19 condition (Long COVID), https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition (2022, accessed 20 March 2023).

[2] Katsarou MS, Iasonidou E, Osarogue A, et al. The Greek Collaborative Long COVID Study: Non-Hospitalized and Hospitalized Patients Share Similar Symptom Patterns. J Pers Med 2022, Vol 12, Page 987 2022; 12: 987.

[3] Pitroda H. Long Covid Sentiment Analysis of Twitter Posts to understand public concerns. 8th Int Conf Adv Comput Commun Syst ICACCS 2022 2022; 140–148.

[4] Kapoteli E, Koukaras P, Tjortjis C. Social Media Sentiment Analysis Related to COVID-19 Vaccines: Case Studies in English and Greek Language. IFIP Adv Inf Commun Technol 2022; 647: 360–372.

[5] Yin H, Song X, Yang S, et al. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web 2022; 25: 1067–1083.

[6] NLTK :: Natural Language Toolkit, https://www.nltk.org/ (accessed 19 March 2023).

[7] GitHub - stopwords-iso/stopwords-el: Greek stopwords collection, https://github.com/stopwords-iso/stopwords-el (accessed 14 January 2023).

[8] spaCy Usage Documentation, https://spacy.io/usage/linguistic-features#lemmatization (accessed 5 February 2023).

[9] Walker R. rwalk/gsdmm: GSDMM: Short text clustering, https://github.com/rwalk/gsdmm (accessed 15 February 2023).

[10] Koutsikakis J, Chalkidis I, Malakasiotis P. GREEK-BERT: The Greeks visiting Sesame Street. In: 11th Hellenic Conference on Artificial Intelligence, pp. 110–117.

[11] Skroutz Sentiment Analysis with BERT (Greek). Kaggle, https://www.kaggle.com/code/nikosfragkis/skroutz-sentiment-analysis-with-bert-greek/notebook (accessed 19 March 2023).